SHANNON INFORMATION AND THE MUTUAL INFORMATION OF TWO RANDOM VARIABLES

MARCELLO DELGADO

Abstract. In his 1948 paper "A mathematical theory of communication", Shannon establishes a method for quantifying information. His focus was on an information source as a Markov process that produces symbols by moving from states to state based on probabilities. Since he believes that a communication system must be ready to handle any incoming message, so a primary concern of the theory was how uncertain the receiver is of the next incoming symbol. To measure this uncertainty, he created a measure he termed "entropy." This measure can then be used to quantify the dependency between two random variables. This measure, termed the mutual information between two random variables, quantifies the amount of information one random variable gives about the other.

Contents

1. Communication Systems and Strings

A communication system consists of an information source, a transmitter, a channel, a receiver, and a destination. Information sources produce messages or a sequence of messages. The transmitter then encodes the message into a signal that is suitable for transmission over the channel, which serves as the medium between the transmitter and the receiver. Assuming all is right with the world, the receiver will have the ability to decode the signal and recover the original message. I will focus here on communication systems that operate using binary strings to encode messages that pass through a noiseless channel.

Definition 1.1 (String)**.** Let $\mathcal{X}$ be some finite or countable set. Then let $\mathcal{X}^*$ denote the set of finite *strings* or sequences over $\mathcal{X}$. Further, let ϵ denote the empty string or word.

Date: DEADLINE AUGUST 22, 2008.

Binary strings are the elements of $\{0,1\}^* = \{\epsilon, 0, 1, 00, 01, 10, 11, 000, \ldots\}$. Each letter of a string is called a *bit*, short for binary digit.

Definition 1.2 (Length)**.** The *length of a string* $x \in \{0,1\}^*$, denoted $l(x)$, is the number of bits in the binary string x.

For example, $l(1010) = 4$ and $l(\epsilon) = 0$. From here we can establish a standard correspondence between the elements of $\mathbb{N}$ and the elements of $\{0,1\}^*$; we will use this correspondence throughout the paper.

$$(0 \to \epsilon), (1 \to 0), (2 \to 1), (3 \to 00), (4 \to 01), (5 \to 10), \ldots$$

If $x \in \{0,1\}^*$ is represented as its corresponding natural number, then $l(x) = \lfloor \log(x+1) \rfloor$. (Throughout this paper, log refers to the logarithm with base of two.)

Definition 1.3 (Probability Mass Function (p.m.f.))**.** Let $\mathcal{A}$ be a finite or countable set, called the sample space. A function $f : \mathcal{A} \to [0,1]$ is a *probability mass function* if $\sum_{a \in \mathcal{A}} f(a) = 1$. If A is a random variable associated with $\mathcal{A}$, then we obtain a probability distribution P over $\mathcal{A}$ by setting $P(A = a) = f(a)$. A subset $\mathcal{A} \subseteq \mathcal{X}$ is called an *event*.

We can use probability mass functions to model communication systems. Every "word" the system is capable of transmitting has a certain probability of being used; this can be modeled by a p.m.f.

2. Codes

2.1. **Preliminaries.** As stated earlier, we are considering messages sent from a transmitter to a receiver. The information being sent is an element of sampl space $\mathcal{A}$, and our system operates by encoding this message into binary strings. Upon receiving the encoded message, the receiver will employ a decoding function, $D : \{0,1\}^* \to A$, to recover the message, where $D(s) = a$ means that "the string s codes for the word a".

Definition 2.1 (Codeword)**.** The elements of subset $D^{-1}(\mathcal{X}) \subseteq \{0,1\}^*$ are called *codewords*.

It is important that this relation be a function so that each received string has only one interpretation. From here, we can intuitively define an encoding relation that is the preimage of the decoding function, $E = D^{-1}$, where for any $x \in \mathcal{A}$, $E(x) = D^{-1}(x) = \{a \in \{0,1\}^* | D(a) = x\}$. E is not necessarily a function.

For each decoding function D, we can define a length function $L_D : \mathcal{A} \to \mathbb{N}$, such that $L_D(x) = \min\{l(y) | D(y) = x\}$. If there exists l_0 s.t. $l_0 = l(y)$ for all $y \in \mathcal{X}$, then D is said to be a fixed-length code. For a fixed-length binary code over a finite set $\mathcal{X}$, $2^{l_0} \geq |\mathcal{X}|$ for each string to to have a unique interpretation, so $l_0 \geq \log |\mathcal{X}|$.

As an example, consider the code:

$$D : \{0,1\}^* \to \{\text{words of the sentence "The quick brown fox jumped over the lazy dog"}\},$$

where $D(0)$ =the, $D(1)$ =quick, $D(00)$ =brown, etc. Now, suppose you recieve the string 0100011011000001010 over the channel and want to decode it to recover the message. You may parse it as

$$D(0)D(1)D(00)D(01)D(10)D(11)D(000)D(001)D(010)$$

and get the intended message "The quick brown fox jumped over the lazy dog". However, you may instead parse the string as

$$D(010)D(00)D(11)D(01)D(10)D(00)D(000)D(10)D(10)$$

to recover the nonsensical message "dog brown over fox jumped brown the jumped jumped".

2.2. **Prefix Codes.** As shown by the previous example, for any given code, it is not always obvious where one codeword ends and the next begins. However, this problem is solved if no codeword is entirely contained in another. In other words, if no allowed string is a *prefix* of another.

Definition 2.2 (Prefix and Prefix-Free)**.** A binary string x is a *proper prefix* of another string y if there exists $z \neq \epsilon$ such that $y = xz$. A set $A \subseteq \{0,1\}^*$ is *prefix-free* if for all pairs of distinct elements of A, neither is a proper prefix of the other. A code $D : A \subseteq \{0,1\}^* \to \mathcal{X}$ is a *prefix code* if A is prefix-free.

We can construct a prefix code if for all strings $x = x_1x_2 \dots x_n$ we let

$$\overline{x} = \underbrace{11\dots1}_{n \text{ times}} 0x_1x_2 \dots x_n.$$

This defines a prefix code, $D : \overline{\{0,1\}^*} \to \mathcal{X}$, since we can tell when a codeword finishes without having to back up. Note, however, that $; (\overline{x}) = 2n+1$, which shows that this code is prefix free at the price of having codewords that are twice as long. We can apply a similar construction applied to $l(x)$ as opposed to x. To do this, let $x' = \overline{l(x)}x$, where $l(x)$ is interpreted using its corresponding string. Then $D(x') = x$ is a prefix code since $(\{0,1\}^*)'$ is prefix-free. Further, for a string $x = x_1x_2 \dots x_n$, $l(x') = n + 2\log n + 1$.

It makes intuitive sense that in a prefix code, shorter codewords are more "expensive": the shorter the codewords used, the fewer such codewords can be used. This notion is formalized in Kraft's Inequality.

Theorem 2.3 (Kraft's Inequality)**.** *Let $\ell_1, \ell_2, \dots, \ell_n$ be a sequence of natural numbers. There is a prefix-code with this sequence as lengths of its binary codewords iff*

$$\sum_i 2^{\ell_i} \leq 1$$

Proof. Any given prefix code can be represented by a binary tree where the root is the empty word ϵ, each branch represents a choice from one of the two "letters" in the binary alphabet $\{0,1\}$, and each codeword i is a path to a node at depth ℓ_i. It follows that the node at which the path of a codeword ends must be a leaf since if it were not then any nodes attached to it would form codewords with prefixes. Then for any leaf in this code tree, let A_i represent the set of descendants that this leaf would have in a full binary tree at a depth of ℓ_n, where ℓ_n is the length of the longest codeword. For $i \neq j$, $A_i \cap A_j = \emptyset$, and further, $|A_i| = 2^{\ell_n - \ell_i}$. Given that the number of nodes at a depth of ℓ_n is 2^{ℓ_n}, then $\sum_i 2^{\ell_n - \ell_i} \leq 2^{\ell_n}$. Dividing both sides by 2^{ℓ_n}, we get the result

$$\sum_i 2^{\ell_i} \leq 1$$

Conversely, given a set of lengths $\ell_1, \ell_2, \ldots, \ell_n$ consider a full binary tree of depth ℓ_n, where ℓ_n is the largest of the ℓ_i. For each l_i, choose a node at depth ℓ_i and remove all of its descendants, □

3. Entropy Intro

When we consider an information source, you can imagine it producing the message symbol by symbol. Each subsequent letter or symbol is chosen according to some probabilities, based both on the probability distribution on the symbols themselves and on the symbols already chosen. A process that carries out actions according to the probabilities of the possible actions is called a *stochastic*, or *Markov*, *process*. We can imagine an information source moving from state to state as it produces its message symbol by symbols, where each of these states has a different set of allowed symbols.

Once we know the probabilities associated with a Markov process, then there should be a way to quantify how uncertain we are of the next symbol the Markov process will produce. Such a measure is motivated from two points of view: the axiomatic approach and the coding approach. For a probability distribution P on $\mathcal{X} = \{1, 2, \ldots, n\}$, $H(\mathcal{X}) = H(P) = H(p_1, p_2, \ldots, p_n)$ denotes the uncertainty, or *entropy*, of the distribution on the finite sample space $\mathcal{X}$. The axiomatic approach postulates that such a measure of uncertainty have certain properties:

(1) H be continuous in the p_i
(2) If for all i, $p_i = \frac{1}{n}$, then H should be a monotonic increasing function of n. The intuitive justification is that as the number of equiprobable events increases, the more uncertain we are of the outcome.
(3) When a choice is broken down into a two-step process, the entropy of P should be equal to the sum of the entropy of the probabilities of the first step of the generation process plus the weighted sum of the entropies in the second step of the generation process. As an example, consider $H(\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$, which is a choice among three. We can instead break this up so that we choose between two choices each with probability $\frac{1}{2}$, $H(\frac{1}{2}, \frac{1}{2})$, and then one of these choices leads to a second choice between two options with probabilities $\frac{1}{3}$ and $\frac{2}{3}$, $H(\frac{1}{3}, \frac{2}{3})$. Then $H(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}) = H(\frac{1}{2}, \frac{1}{2}) + \frac{1}{2}H(\frac{1}{3}, \frac{2}{3})$.

The only function that satisfies all of these criteria is $H(\mathcal{X}) = \sum_{i=1}^{n} p_i \log \frac{1}{p_i}$.

Definition 3.1. The function $H(P) = H(\mathcal{X}) = \sum_{i=1}^{n} p_i \log \frac{1}{p_i}$ describes the *entropy* of the probability distribution P over the finite sample space $\mathcal{X}$.

For an information source with the probability distribution P over its set of symbols, H is a measure of how *uncertain* we are of the outcome of the Markov process operating under P. This is equivalent to the average amount of information we gain by observing an outcome of this Markov process. For a sample space $\mathcal{X}$, H maximizes when the probability distribution over $\mathcal{X}$ is the uniform distribution, which makes intuitive sense since we are most uncertain of the outcome when every outcome is equally likely.

Theorem 3.2. *Let $\mathcal{X}$ be finite s.t. $|\mathcal{X}| = n$. Then for a uniform probability distribution over $\mathcal{X}$, $H(\mathcal{X}) = \log n$. Moreover,, the uniform distribution maximizes $H(\mathcal{X})$.*

Proof. Consider the uniform distribution over the elements of $\mathcal{X}$. Then

$$H(\mathcal{X}) = \sum_{i=1}^{n} \frac{1}{n} \log n = n(\frac{1}{n} \log n) = \log n.$$

Claim 3.3. *For all x, $\log x \leq x - 1$ and equality holds iff $x = 1$.*

By definition, $\log x = \int_1^x \frac{1}{t} dt$. Further, $x = \int_1^x dt + 1$, which means $x - 1 = \int_1^x dt$. Consider $x > 1$. The function $\frac{1}{t}$ is decreasing function since for all $t \geq 1$, $\frac{1}{t} \leq 1$. This implies that $\int_1^x \frac{1}{t} dt \leq \int_1^x dt$ for $x \geq 1$ since $\int_1^x \frac{1}{t} dt > \int_1^x dt$ would require there to be $t \geq 1$ for which $\frac{1}{t} > 1$, which is impossible. For $0 < t < 1$, then $\frac{1}{t} > 1$ for all such t. Thus, $\int_x^1 \frac{1}{t} dt > \int_x^1 dt$. This implies that $\int_1^x \frac{1}{t} dt < \int_1^x dt$ for all such x. Therefore, for all positive x, $\int_1^x \frac{1}{t} dt \leq \int_1^x dt$, which implies that $\log x \leq x - 1$.

If $x = 1$, then $\int_1^1 \frac{1}{t} dt = \int_1^x dt = 0$. If $\log x = x - 1$, then $\log x - x + 1 = 0$. $x = 1$ is a solution since $\log 1 - 1 + 1 = 0$. Assume there exists another solution y such that $\log y - y + 1 = 0$ and $y \neq 1$. $y > 0$ since the log function is undefined for $x \leq 0$. If $y > 1$, then by Rolle's theorem, there exists $z \in (1, y)$ s.t. $\frac{1}{z} - 1 = 0$. However, this implies that $z = 1$, which is a contradiction since z must be between 1 and y. An analogous proof shows that no such $0 < y < 1$ can exist. Thus, if $\log x = x - 1$, then $x = 1$.

Claim 3.4. *For a probability distribution P and a (sub)probability distribution Q over n events (that is, $\sum_{i=1}^n p_i = 1$ and $\sum_{i=1}^n q_i \leq 1$) then $\sum_{i=1}^n p_i \log \frac{1}{p_i} \leq \sum_{i=}^n p_i \log \frac{1}{q_i}$ and equality holds iff $p_i = q_i$ for all i.*

By the properties of logarithms, $\log \frac{q_i}{p_i} = \log q_i - \log p_i$ for any i. Using the above claim, $\log q_i p_i \leq \frac{q_i}{p_i} - 1$ for all i. Transitively, $\log q_i - \log p_i \leq \frac{q_i}{p_i} - 1$, which implies that $p_i \log q_i - p_i \log p_i \leq q_i - p_i$. By properties of logarithms, $p_i \log q_i - p_i \log p_i = p_i \log \frac{1}{p_i} - p_i \log \frac{1}{q_i}$. It follows that

$$\begin{aligned} p_i \log \frac{1}{p_i} - p_i \log \frac{1}{q_i} &\leq q_i - p_i, \text{ for all } i, \text{ therefore} \\ \sum_{i=1}^{n} p_i \log \frac{1}{p_i} - \sum_{i=1}^{n} p_i \log \frac{1}{q_i} &\leq \sum_{i=1}^{n} q_i - \sum_{i=1}^{n} p_i \leq 0 \end{aligned}$$

Thefore, $\sum_{i=1}^n p_i \log \frac{1}{p_i} \leq \sum_{i=1}^n p_i \log \frac{1}{q_i}$.

For all i, $p_i \log \frac{1}{p_i} - p_i \log \frac{1}{q_i} \leq q_i - p_i$ follows bidirectionally from $\log \frac{q_i}{p_i} \leq \frac{q_i}{p_i} - 1$. Then for all i, $\log \frac{q_i}{p_i} = \frac{q_i}{p_i} - 1$ iff $\frac{q_i}{p_i} = 1 \Rightarrow q_i = p_i$. Thus for all i, $p_i \log \frac{1}{p_i} - p_i \log \frac{1}{q_i} = q_i - p_i = 0$ iff $p_i = q_i$. And from there it follows that $\sum_{i=1}^n p_i \log \frac{1}{p_i} - \sum_{i=1}^n p_i \log \frac{1}{q_i} = 0$, so therefore $\sum_{i=1}^n p_i \log \frac{1}{p_i} = \sum_{i=1}^n p_i \log \frac{1}{q_i}$ iff $p_i = q_i$ for all i.

Claim 3.5. $\sum_{i=1}^n p_i \log \frac{1}{p_i} \leq \log n$.

By previous claim, for any two probability distributions P and Q, $\sum_{i=1}^n p_i \log \frac{1}{p_i} \leq \sum_{i=}^n p_i \log \frac{1}{q_i}$ and equality holds iff $p_i = q_i$ for all i. If Q is the uniform distribution, then $\sum_{i=1}^n p_i \log \frac{1}{p_i} \leq \sum_{i=}^n p_i \log n$.

$$\begin{aligned}\sum_{i=}^{n} p_i \log n &= (\log n)\sum_{i=1}^{n} p_i \\ &= \log n\end{aligned}$$

Therefore, for any probability distribution P, $\sum_{i=1}^{n} p_i \log \frac{1}{p_i} \leq \log n$. Further, the previous claim states that equality holds iff $p_i = q_i$ for all i. So $\sum_{i=1}^{n} p_i \log \frac{1}{p_i} = \log n$ iff P is also the uniform distribution. Therefore, the uniform distribution maximizes entropy in the finite case. □

4. Mutual Information

Consider two sample spaces $\mathcal{X}, \mathcal{Y}$ with associated random variables X and Y. How much does an outcome $X = x$ tell us about which value Y will take? Let U be the set of possible events $(X = x, Y = y)$ that consist of the joint occurrence of events $X = x$ and $Y = y$. If U is not the whole set $\mathcal{X} \times \mathcal{Y}$, then there must be some dependency between X and Y. We can extend the definition of entropy from one random variable to a joint entropy for two.

Definition 4.1. Let $H(X, Y)$ denote the *joint entropy of random variables X and Y*. If f is a joint probability mass function, where $f(x, y)$ denotes the probability of the event $(X = x, Y = y)$, then

$$H(X, Y) = \sum_{x,y} f(x, y) \log \frac{1}{f(x, y)}$$

When dealing with two random variables, then we can define the probabilities of an event in one random variable in terms of the other.

Definition 4.2. We can define the probability of an event $X = x$ occurring in the following way:

$$f_1(x) = \sum_{y} (x, y)$$

and the probability that an event $Y = y$ occurs in the following way:

$$f_2(y) = \sum_{x} (x, y)$$

From this, we can extend to a definition of conditional probability.

Definition 4.3. For events $X = x$, $Y = y$, the *conditional probability* of $Y = y$ given $X = x$, denoted $f(y|x)$, is given by

$$f(y|x) = \frac{f(x, y)}{\sum_y f(x, y)} = \frac{f(x, y)}{f_1(x)},$$

and analogously, the conditional probability of $X = x$ given that $Y = y$, denoted $f(x|y)$, is given by

$$f(x|y) = \frac{f(x, y)}{\sum_x f(x, y)} = \frac{f(x, y)}{f_2(y)}.$$

Further, two events are independent if either $f(y|x) = f_2(y)$ or $f(x|y) = f_1(x)$, and thus $f(x, y) = f_1(x)f_2(y)$.

We can also define conditional entropy, which measures the uncertainty in one random variable once we have knowledge of the other.

Definition 4.4 (Conditional Entropy)**.** For random variables X and Y, the *conditional entropy of Y given that $X = x$* is defined as

$$H(Y|X = x) = \sum_y f(y|x) \log \frac{1}{f(y|x)}$$

Further, we can define the *conditional entropy of Y given X* as the average of the conditional entropies of Y for each possible outcome of X weighted by the probability of that outcome. Formally,

$$H(Y|X) = \sum_x f_1(x) H(Y|X = x)$$

It makes intuitive sense that the joint entropy of two random variables should not exceed the sum of their individual entropies, and we can check this.

Lemma 4.5. *Suppose we reformulate the definition of entropies of X and Y so that $H(X) = \sum_x f_1(x) \log \frac{1}{f_1(x)}$ and $H(Y) = \sum_y f_2(y) \log \frac{1}{f_2(y)}$. Then $H(X, Y) \leq H(X) + H(Y)$ and equality holds if X and Y are independent.*

Proof. We know by definition that $H(X, Y) = \sum_{x,y} f(x, y) \log \frac{1}{f(x,y)}$. Since summing over all pairs x and y is equivalent to summing over x and then summing over y. Therefore, $H(X, Y) = \sum_y \sum_x f(x, y) \log \frac{1}{f(x,y)}$.

$$\begin{aligned}
H(X, Y) &= \sum_y \sum_x f(x, y) \log \frac{1}{f(x, y)} \\
&= \sum_y \sum_x f_2(y) f(x|y) \log \left(\frac{1}{f_2(y) f(x|y)} \right) \\
&= \sum_y f_2(y) \sum_x f(x|y) \left(\log \frac{1}{f_2(y)} + \log \frac{1}{f(x|y)} \right) \\
&= \sum_y f_2(y) \log \frac{1}{f_2(y)} \sum_x f(x|y) + \sum_y f_2(y) \sum_x f(x|y) \log \frac{1}{f(x|y)} \\
&= \sum_y f_2(y) \log \frac{1}{f_2(y)} \sum_x \frac{f(x, y)}{f_2(y)} + \sum_y f_2(y) H(X|Y = y)
\end{aligned}$$

Since $f_2(y) = \sum_x f(x, y)$, then $\sum_x \frac{f(x,y)}{f_2(y)} = \frac{1}{f_2(y)} \sum_x f(x, y) = 1$. Further, $H(X|Y) = \sum_y f_2(y) H(X|Y = y)$, so therefore

$$\sum_y f_2(y) \log \frac{1}{f_2(y)} \sum_x \frac{f(x, y)}{f_2(y)} + \sum_y f_2(y) H(X|Y = y) = H(Y) + H(X|Y)$$

Therefore, $H(X, Y) = H(Y) + H(X|Y)$. We know that $H(X|Y) \leq H(X)$ since the knowledge of Y can never increase our uncertainty of the outcome of X. Further, $H(X) = H(X|Y)$ iff X and Y are independent. Therefore,

$$H(Y) + H(X|Y) \leq H(Y) + H(X)$$

and so

$$H(X, Y) \leq H(X) + H(Y)$$

and equality holds if and only if X and Y are independent. □

If these two random variables are dependent in any way, then knowledge of one can give us knowledge about the outcome of the other. We can quantify this using the concept of *information*.

Definition 4.6 (Information)**.** For two random variables X and Y, the *information* in the outcome $X = x$ about Y is defined as

$$I(X = x : Y) = H(Y) - H(Y|X = x)$$

This definition makes intuitive sense since it measures the change in the uncertainty about Y that occurs with the knowledge of the outcome $X = x$. From here, we can also see that $I(X = x : X) = H(X)$ since $H(X|X = x) = 0$ since all uncertainty is removed. We can see that $I(X = x : Y)$ and $I(Y = y : X)$ are not directly comparable, but we can however directly compare their expected values.

$$E(I(X = x : Y)) = \sum_x f_1(x) I(X = x : Y)$$

$$E(I(Y = y : X)) = \sum_y f_2(y) I(Y = y : X)$$

Claim 4.7. $E(I(X = x : Y)) = E(I(Y = y : X))$

Proof. We know $E(I(X = x : Y)) = \sum_x f_1(x) I(X = x : Y)$ and that $I(X = x : Y) = H(Y) - H(Y|X = x)$. Therefore,

$$E(I(X = x : Y)) = \sum_x f_1(x)(H(Y) - H(Y|X = x))$$

$$\begin{aligned} \sum_x f_1(x)(H(Y) - H(Y|X = x)) &= H(Y) \sum_x f_1(x) - \sum_x f_1(x) H(Y|X = x) \\ &= H(Y) - H(Y|X) \end{aligned}$$

We know $H(X, Y) = H(X) + H(Y|X)$, which implies $H(X, Y) - H(X) = H(Y|X)$. Therefore,

$$H(Y) - H(Y|X) = H(Y) - (H(X, Y) - H(X)) = H(X) + H(Y) - H(X, Y)$$

We also know $E(I(Y = y : X)) = \sum_y f_2(y) I(Y = y : X)$ and that $I(Y = y : X) = H(X) - H(X|Y = y)$. Therefore,

$$\begin{aligned} E(I(X = x : Y)) &= \sum_y f_2(y)(H(X) - H(X|Y = y)) \\ &= H(X) \sum_y f_2(y) - \sum_y f_2(y) H(X|Y = y) \\ &= H(X) - H(X|Y) \\ &= H(X) - (H(X, Y) - H(Y)) \\ &= H(X) + H(Y) - H(X, Y) \end{aligned}$$

Therefore, by transitivity $E(I(X = x : Y)) = E(I(Y = y : X))$. □

Definition 4.8 (Mutual Information)**.** We define the *mutual information between random variables X and Y* as the common value

$$I(X;Y) = E(I(X = x : Y)) = E(I(Y = y : X))$$

. Further,

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= \sum_x \sum_y f(x,y) \log \frac{f(x,y)}{f_1(x) f_2(y)} \end{aligned}$$

References

[1] C. E. Shannon A Mathematical Theory of Communication, 1948
[2] P Grunwald and P Vitanyi Shannon Information and Kolmogorov Complexity, 2008